DOCUMENT RESUME

ED 467 377 TM 034 303

AUTHOR	Glas, Cees A. W.; van der Linden, Wim J.
TITLE	Modeling Variability in Item Parameters in Item Response Models. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
SPONS AGENCY	Law School Admissions Council, Newtown, PA.
REPORT NO	RR-01-11
PUB DATE	2001-00-00
NOTE	41p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Bayesian Statistics; *Estimation (Mathematics); Item Response Theory; Markov Processes; *Models; Monte Carlo Methods
IDENTIFIERS	Gibbs Sampling; *Item Parameters; *Variability

ABSTRACT

In some areas of measurement item parameters should not be modeled as fixed but as random. Examples of such areas are: item sampling, computerized item generation, measurement with substantial estimation error in the item parameter estimates, and grouping of items under a common stimulus or in a common context. A hierarchical version of the three-parameter normal ogive model is used to model parameter variability in multiple populations of items. Two Bayesian procedures for the estimation of the parameter are given. The first method produces an estimate of the posterior distribution using a Markov Chain Monte Carlo method (Gibbs sampler); the second procedure produces a Bayes modal estimate. It is shown that the procedure using the Gibbs sampler breaks down if for some of the random item parameters the sampling design yields only one response. However, in this case, marginalization over the item parameters does result in a feasible estimation procedure. Some numerical examples are given. (Contains 2 tables, 4 figures, and 36 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 467 377

TM

Modeling Variability in Item Parameters in Item Response Models

Research Report 01-11

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Cees A.W. Glas
W.J. van der Linden

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM034303

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

BEST COPY AVAILABLE

Modeling Variability in Item Parameters In Item Response Models

Cees A.W. Glas

Wim J. van der Linden

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this paper are those of the author and do not necessarily reflect the position or policy of LSAC. We thank the Ngee Ann Polytechnic in Singapore and Cito in the Netherlands for use of their data. Requests for information should be send to:
Cees A.W. Glas, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
E-mail: C.A.W.Glas@edte.utwente.nl.

Abstract

In some areas of measurement item parameters should not be modeled as fixed but as random. Examples of such areas are: item sampling, computerized item generation, measurement with substantial estimation error in the item parameter estimates, and grouping of items under a common stimulus or in a common context. A hierarchical version of the three-parameter normal-ogive model is used to model parameter variability in multiple populations of items. Two Bayesian procedures for the estimation of the parameter are given. The first method produces an estimate of the posterior distribution using a Markov chain Monte Carlo method (Gibbs sampler), the second produces a Bayes modal estimate. It is shown that the procedure using the Gibbs sampler breaks down if for some of the random item parameters the sampling design yields only one response. However, in this case, marginalization over the item parameters does result in a feasible estimation procedure. Some numerical examples are given.

Keywords: Bayesian estimates; Bayes modal estimates; Gibbs sampler; item generation; item grouping; item sampling; multilevel item response theory; marginal maximum likelihood; Markov chain Monte Carlo; sampling design.

Introduction

Item response theory (IRT) models with random examinee parameters have become a common choice among practitioners in the field of educational measurement. Though initially the choice for such models was motivated by the attempt to get rid of the statistical problems inherent in the incidental nature of the examinee parameters (Bock & Lieberman, 1970), the insight soon emerged that such models more adequately represent cases where the focus is not on measurement of individual examinees but estimation of characteristics of populations. Early examples of models with random examinee parameters in the literature are given in Andersen and Madsen (1977) and Sanathanan and Blumenthal (1978), who were interested in estimates of the mean and variance in a population of examinees, and in Mislevy (1991), who provided the tools for inference from a response model with a regression structure on the examinee parameters introduced to account for sampling from populations of examinees with different values on background variables.

In traditional large-scale testing, a statistical necessity to model *item* parameters in IRT models as random has hardly been felt. Typically, the values of the item parameters are first estimated from large samples of examinees, with the examinee parameter integrated out of the likelihood or posterior distribution. During operational testing, because the calibration sample was large, the item parameters are fixed to their estimates and treated as known constants rather than as incidental parameters with unknown values. Nevertheless, the measurement literature shows a recent interest in response models with random item parameters. The reason for this phenomenon is the insight that such models better represent the use of sampling designs that involve random selection of items or cases where sets of items can be considered as exchangeable once we know they belong to the same "group" or "class".

The most obvious case of measurement with random item characteristics arises is domain-referenced testing. In this type of testing, the idea of assembling a fixed test for all examinees is abandoned in favor of a random sample from a large pre-written pool of items for each examinee (e.g., Millman, 1973). The model originally used to guide

domain-referenced testing programs with dichotomously scored items was the binomial error model (Lord & Novick, 1968, chap. 23), given by

$$\Pr\{X_n = x \mid k, \pi_i\} = \binom{k}{x} \pi_n^x (1 - \pi_n)^{k-x},$$

where X_n is the number of successes for examinee n on a test of size k sampled from the domain and π_n is the examinee's success parameter. Clearly, the success parameter in this model depends both on the examinee and the domain of test items. Attempts to decompose π_n into separate components for the examinee and the items led to the introduction of IRT models with random item parameters. One of the first models of this kind is found in Albers, Does, Imbos and Jansen (1989), who needed an explicit examinee parameter to estimate progress of learning in a longitudinal study with tests sampled from the same pool of items at different time points.

A more sophisticated application of the idea of item sampling has become available through the introduction of computer-generated items in educational measurement. Using an item-cloning technique (see, for instance, Bejar, 1993, or Roid & Haladyna, 1982), it is no longer necessary to write each item in the domain individually. Instead they can be generated by the computer from a smaller set of "parent items" through the use of transformation rules. One of the more popular types of computer generation of items is based on so-called "replacement set procedures" (Millman & Westman, 1989), where the computer is used to replace elements in the parent item (e.g., key terms, relations, numbers, and distractors) randomly from well-defined sets of alternatives. Because the substitution introduces (slight) random variation between items derived from the same parent, it becomes efficient to model the item parameters as random and shift the interest to the hyperparameters that describe the distributions of the item parameters within parents (Glas & van der Linden, 2001). Observe that this application is more general than the previous one because we now consider sampling from multiple populations of items in the same test.

The current trends towards increased testing in education and individualization of test administration have put stress on the resources for item calibration at testing organizations. As a consequence, it becomes attractive to find alternatives to the traditional large-sample approach to item calibration. A possible solution is to accept non-negligible estimation error in item parameter estimates and treat them as random in operational testing, e.g., using their posterior distribution when assembling a test or estimating examinee parameters. The first to deal with this problem in IRT were Tsutakawa and Johnson (1990; see also van der Linden & Pashley, 2000). The problem of how to deal with posterior distributions for the item parameters in an adaptive testing procedure has been addressed in Glas and van der Linden (2001).

An omnipresent feature of mainstream IRT models is the assumption of conditional independence between the response variables given the examinee's ability level. However, it has long been known that items that share a common element may loose this feature. Examples are sets of items with a common stem or items sharing a common context because the test is organized as a set of fixed testlets (Wainer & Kiely, 1987). To deal with this problem, Bradlow, Wainer and Wang (1999; see also Wainer, Bradlow & Du, 2001) replaced the well-known parameter structure in two-parameter and three-parameter IRT models by

$$a_i(\theta_n - b_i - \gamma_{nd(i)}),$$

where θ_n, b_i and a_i are the traditional parameters for the ability of examinee n and the difficulty and discrimination power of item i. The new parameter $\gamma_{nd(i)}$ was introduced to represents a random effect for the combination of examinee n and the nesting of item i in testlet d. Observe that this model actually is an (overparameterized) version of a multidimensional IRT model with decomposition $\gamma_{nd(i)} = a_{id}\theta_{dn}$, where θ_{dn} is the score for examinee n on an ability dimension unique to testlet d and a_{id} is the discrimination parameter for item i on this dimension. Because testlets have a fixed structure, randomness of $\gamma_{nd(i)}$ cannot come from sampling of the items. However, if the examinees are sampled, θ_{dn} becomes random, and so does $\gamma_{nd(i)}$.

A final example of the use of a model with random item parameters is given in Janssen, Tuerlinckx, Meulders and de Boeck (2000). These authors are interested in the process of standard setting on a criterion-referenced test with sections of items in the test grouped under different criteria. Because of this grouping, the IRT model is chosen to have random item parameters with different distributions for different sections. At first sight, grouping of items does not necessarily seem to lead to a model with random parameters. However, a general approach to account for dependency due to common elements between units is to behave as if they were a stratified random sample from a set of subpopulations and model the process accordingly. A Bayesian argument in favor of this approach is that if the only thing known a priori about the items is that they are grouped under common criteria, they are exchangeable given the criterion and can be treated as if they are a random sample.

It is the purpose of this article to give a Bayesian treatment of the problem of estimating the parameters in a model with random item parameters and multiple populations of items. The model does not only allow for all item properties that have traditionally been modeled using the three-parameter logistic model (item difficulty, discriminating power, and possibility to guess) but also for dependency between these features within populations (e.g., correlation between parameters for discriminating power and guessing). The treatment is fully Bayesian in the sense that (informative) priors are formulated for all hyperparameters describing the distributions of the item parameters within the populations. Two estimation procedures are presented. In the first procedure, the posterior distribution of all parameters are generated concurrently using a Markov chain Monte Carlo (MCMC) simulation algorithm (i.e., the Gibbs sampler). In the second procedure, Bayesian modal estimates for a subset of the parameters are computed marginalizing over the other parameters. Before presenting the procedures, a feature of the sampling design for collecting the response data critical for the choice between the parameter estimation procedures is discussed.

Sampling Design

The sampling design governs sampling of items and examinees in the calibration study and thus controls how much response data we have for each possible realization of the random item and examinee parameters. A critical feature for item parameter estimation in the multilevel model below is the number of responses per realization of the random item parameters. If, as will become clear below, this number is equal to one for some of the items, a procedure for concurrently estimating the posterior distributions of all parameters in the model, breaks down in the sense that we have too little data, that is, no statistical information can be aggregated for the some of the parameters in the model. In the sequel, these item parameters will be called incidental item parameters.

A practical illustration of the distinction between a sampling design where all the item parameters can be treated as structural and a sampling design where some of the item parameters are incidental is the case of computer-generated items discussed above. One possible implementation of computer-based item generation is to have the computer generate a new item for each examinee ("item generation on the fly"). Another implementation is to generate a set of item clones prior to operational testing and sample from this set during testing. In the former case, all item parameters are incidental; in the latter case, some items will have incidental parameters if the set is large relative to the population of examinees tested and the design involves random assignment of items to examinees (as, for instance in adaptive testing).

The distinction between structural and incidental parameters in statistical models has been introduced by Neyman and Scott (1948; also see, Kiefer & Wolfowitz, 1956). In an estimation problem with structural parameters, the number of parameters remains finite if the number of observations goes to infinity, whereas in a problem with incidental parameters the number of parameters goes to infinity. The presence of incidental parameters causes problems for statistical inferences, for instance, the solutions to the likelihood equations for the structural parameters may loose their consistency or asymptotic efficiency.

If each examinee gets a different item, the random item parameters are incidental parameters in the sense of Neyman and Scott. If the items are sampled from a finite set, their parameters are structural. However, the latter may still result in inestimable parameters in the Bayesian framework below. Nevertheless, one of the proven measures to solve problems with incidental parameters – marginalizing them out of the likelihood function – also works for the case in which some examinees get unique items. For such cases a marginal maximum likelihood approach is presented. This solution can be used as an alternative to the Bayesian framework in testing with computerized item generation if new items are generated on the fly for each examinee, or in any other application of response models with random item parameter with too few responses per item.

We will return to this issue in the Discussion section to discuss other sampling designs that complicate parameter estimation in models with random item parameters. In fact, as already admitted in Newman and Scott (1948), more complex cases exist in which parameters appear in varying combinations of random variables. Educational measurement with random item parameters and incomplete sampling design clearly belongs to this category.

The Model

Consider a set of item populations $p = 1, ..., P$ of size $k_1, ..., k_P$, respectively. The items in population p will be labeled $i_p = 1, ..., k_p$. It proves convenient to introduce sampling design variables d_{ni_p}, which assumes a value equal to one if person n responded to item i_p and zero otherwise. Let X_{ni_p} be the response variable for person n and item i_p. If $d_{ni_p} = 1$, X_{ni_p} attains the value one for a correct response and a value zero for an incorrect response. If $d_{ni_p} = 0$, X_{ni_p} attains an arbitrary value r ($r \neq 0; r \neq 1$). Notice that with this definition the design variables are completely determined by the response variables; they are only introduced to facilitate the mathematical presentation.

First-Level Model

The first-level model is the three-parameter normal ogive (3PNO) model, which describes the probability of a correct response as

$$p(x_{ni_p} = 1 \mid d_{ni_p} = 1, \theta_n, a_{i_p}, b_{i_p}, c_{i_p}) = c_{i_p} + (1 - c_{i_p})\Phi(a_{i_p}\theta_n - b_{i_p}), \qquad (1)$$

where a_{i_p}, b_{i_p},and c_{i_p} are item parameters, θ_n is an examinee parameter, and $\Phi(.)$ is the normal cumulative distribution function. The parameterization of the models in (1) is slightly different from the usual parameterization for the logistic and normal-ogive models, $a_{i_p}(\theta - b_{i_p})$. The only motivation for our choice is to simplify the presentation below.

The reason for considering the 3PNO model rather that the 3PL model is that the former appears to be more tractable in an MCMC framework. However, as is well known, for an appropriately chosen scale factor both models are numerically nearly indistinguishable and either model is expected to fit only if the other does.

Second-Level Model

The values of the item parameters $(a_{i_p}, b_{i_p}, c_{i_p})$ in (1) are considered as realizations of a random vector. We will use the transformation

$$\xi_{i_p} = (a_{i_p}, b_{i_p}, \text{logit } c_{i_p}), \qquad (2)$$

which gives the item parameters scales for which the following assumption of multivariate normality is reasonable:

$$\xi_{i_p} \sim N(\mu_p, \Sigma_p), \qquad (3)$$

where μ_p is the vector with the mean values of the item parameters for population p and Σ_p their covariance matrix. Observe that the hyperparameters (μ_p, Σ_p) are allowed to vary across the populations of items.

In the inferences below, we assume that θ_n has a standard normal distribution

$$\theta_n \sim N(0,1). \tag{4}$$

This assumption holds if examinee n is from a population of exchangeable examinees with a normal distribution of abilities. Examinees and items are thus distributed independently, that is, we do not assume that the items are sampled dependently on the examinee abilities.

Prior for Hyperparameters

A convenient choice for the prior distribution for the hyperparameters (μ_p,Σ_p) is a normal-inverse-Wishart distribution (see, for instance, Box & Tiao, 1973, or Gelman, Carlin, Stearn & Hall, 1995). The prior follows from the specification

$$\Sigma_p \sim Inv - Wishart_{v_0}(\Sigma_0)$$

$$\mu_p \mid \Sigma_p \sim MVN(\mu_0, \Sigma_p/\kappa_0)$$

and has a density given by

$$p(\mu_p, \Sigma_p) \propto |\Sigma_0|^{-((v_0+3)/2+1)} \exp\left(-\frac{1}{2}tr(\Sigma_0\Sigma_p^{-1}) - \frac{\kappa_0}{2}(\mu_p-\mu_0)^T\Sigma_p^{-1}(\mu_p-\mu_0)\right), \tag{5}$$

where Σ_0 and v_0 are the scale matrix and degrees of freedom for the prior on Σ_p and μ_0 and κ_0 are the weight for the prior on μ_p, respectively. The weight expresses the information in the prior distribution as the number of prior measurements it can be equated to.

It should be noted that, though the hyperparameters (μ_p,Σ_p) are allowed to take different values across populations, a common prior is specified for all hyperparameters. The function of the prior is only to bound their distribution to a likely region of possible values.

Likelihood Function

The response vector of examinee n is denoted as $\mathbf{x}_n = (x_{ni_1}, \ldots, x_{ni_p}, \ldots, x_{ni_P})$. Using the assumptions of (1) independence between examinees, (2) independence between items and examinees, and (3) local independence within examinees, the likelihood function associated with response data $\mathbf{x} \equiv (\mathbf{x}_n)$ can be written as

$$
\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x}, (\mathbf{d}_n)) &\equiv \prod_n p(\mathbf{x}_n \mid \mathbf{d}_n, \theta_n, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \prod_n \prod_p \prod_{i_p} p(x_{ni_p} \mid d_{ni_p}, \theta_n, \boldsymbol{\xi}_{i_p}) p(\theta_n) \\
&\quad \prod_p \prod_{i_p} p(\boldsymbol{\xi}_{i_p} \mid \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p). \qquad (6)
\end{aligned}
$$

The convention will be followed that $p(x_{ni_p} \mid d_{ni_p} = 0, \theta_n, a_{i_p}, b_{i_p}, c_{i_p}) = 1$.

Discussion

The current model for random items and multiple item populations differs from the multilevel IRT models for testlets in Bradlow, Wainer & Wang (1999) and Wainer, Bradlow, and Zu (2000) in that the latter only has a random interaction parameter between examinees and items but fixed parameters a_i, b_i, and c_i. The statistical treatment of the models is the same, however; in these two papers an MCMC framework is used to estimate the parameters as well. The current model also differs from the one in Albers, Does, Imbos and Jansen (1989). These authors use a one-parameter version of the normal-ogive model, i.e., the model in (1) with $a_i = 1$ and $c_i = 0$, but add a growth parameter for each examinee that is assumed to increase linearly over time. Finally, the model introduced in Janssen, Tuerlinckx, Meulders and de Boeck (2000) is a two-parameter version of the one in (1) obtained by setting $c_i = 0$. Their second-level model specifies independent normal distributions for a_i and b_i and is thus a special case of (3) with $\boldsymbol{\Sigma}_p$ reduced to a 2x2 identity matrix. These authors also treat parameter estimation in an MCMC framework, but with uninformative priors for (μ_a, μ_b) rather than the prior in (5).

Parameter Estimation

Three methods for estimation of the parameters of the model will be discussed. The first two pertain to sampling schemes where the item parameters ξ_{i_p} can be viewed as structural parameters, that is, as the sample size grows, their number remains limited; or, in other words, the sampling design is such that statistical information with respect to these parameters can be accumulated. The first method is a Bayesian method where the joint posterior distribution of all model parameters is evaluated using the Gibbs sampler. The second method is a Bayes modal estimation procedure that produces point estimates of the item parameters. From a Bayesian perspective, the latter method produces posterior mode estimates of the item parameters ξ_{i_p}, μ_p and Σ_p, where the posterior is marginalized over the incidental parameters θ. The third estimation procedure pertains the case where the item parameters ξ_{i_p} are also incidental. The third procedure is a Bayes modal estimation procedure where the likelihood or the posterior is marginalized both with respect to ξ_{i_p} and θ.

Bayesian Estimation Using the Gibbs Sampler

In Bayesian modeling, all parameters are considered as random variables. A modern approach to produce the posterior joint distribution of the parameters of interest is by simulation. A Markov chain Monte Carlo (MCMC) procedure will be used to sample this posterior distribution. The chains will be constructed using the Gibbs sampler (Gelfand & Smith, 1990). To implement the Gibbs sampler, the parameter vector is divided into a number of components, and the components are sampled consecutively from their conditional posterior distributions given the last sampled values for all other components. This sampling scheme is repeated until the distribution of sampled values forms a stable estimate of the posterior distributions. Albert (1992) applies Gibbs sampling to estimate the parameters of the 2PNO model. A generalization to the 3PNO model is given by Béguin and Glas (2001). A more general introduction to MCMC for IRT models is found in Patz and Junker (1999a), whereas applications for models with multiple raters, multiple item types and missing data are given in Patz and Junker (1999b), models with a multi-

level structure on the ability parameters in Fox and Glas (2001) and multidimensional models in Béguin and Glas (2001) and Shi and Lee (1998).

Data Augmentation

Béguin and Glas (2001) introduce a data augmentation scheme for the 3PNO based on the following interpretation. (In their implementation of the Gibbs sampler they choose a Beta prior for c_{i_p} and a uniform prior on the positive real line for a_{i_p}, though.) Suppose that the examinee knows the correct answer with probability $\Phi(\lambda_{ni_p})$, with $\lambda_{ni_p} = a_{i_p}\theta_n - b_{i_p}$, and then gives a correct response with probability one or does not know the correct answer with probability $1-\Phi(\lambda_{ni_p})$ and then guesses the correct response with probability c_{i_p}. The marginal probability of a correct response is equal to $\Phi(\lambda_{ni_p})+ c_{i_p}(1 - \Phi(\lambda_{ni_p}))$. Let

$$W_{ni_p} = \begin{cases} 1 & \text{if person } i \text{ knows the correct answer to item } j \\ 0 & \text{if person } i \text{ doesn't know the correct answer to item } j. \end{cases} \tag{7}$$

So if $W_{ni_p} = 0$, person i will guess the response to item j, and if $W_{ni_p} = 1$, person i will know the right answer and will give a correct response. Consequently, the conditional probability of $W_{ni_p} = w_{ni_p}$ given $X_{ni_p} = x_{ni_p}$ is given by

$$\begin{aligned} &P(W_{ni_p} = 1 \mid X_{ni_p} = 1, \lambda_{ni_p}, c_{i_p}) \propto \Phi(\lambda_{ni_p}) \\ &P(W_{ni_p} = 0 \mid X_{ni_p} = 1, \lambda_{ni_p}, c_{i_p}) \propto c_{i_p}(1 - \Phi(\lambda_{ni_p})) \\ &P(W_{ni_p} = 1 \mid X_{ni_p} = 0, \lambda_{ni_p}, c_{i_p}) = 0 \\ &P(W_{ni_p} = 0 \mid X_{ni_p} = 0, \lambda_{ni_p}, c_{i_p}) = 1. \end{aligned} \tag{8}$$

In addition to W_{ni_p}, following Albert (1992), the data are also augmented with latent data Z_{ni_p}, which are independent and normally distributed with mean $\lambda_{ni_p} = a_{i_p}\theta - b_{i_p}$ and standard deviation equal to one. The observed data X_{ni_p} are considered as indicators of the sign of Z_{ni_p}; if $X_{ni_p} = 0$ or 1, Z_{ni_p} is negative or positive, respectively.

Posterior Distribution

The aim of the procedure is to simulate samples from the joint posterior distribution given by

$$p(\boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{z}, \mathbf{w} \mid \mathbf{x}) \quad \propto \quad p(\mathbf{z}, \mathbf{w} \mid \mathbf{x}; \boldsymbol{\xi}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\xi} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{9}$$

The right-hand side probability (density) functions are given by (10) (see below), (4), (3) and (2), respectively.

Steps in the Gibbs Sampler

The steps of the Gibbs sampler are the following.

Step 1

The posterior $p(\mathbf{z}, \mathbf{w} \mid \mathbf{x}; \boldsymbol{\xi}, \boldsymbol{\theta})$ is factored as $p(\mathbf{z} \mid \mathbf{x}; \mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\theta}) p(\mathbf{w} \mid \mathbf{x}; \boldsymbol{\xi}, \boldsymbol{\theta})$. For the cases with $d_{ni_p} = 1$, the values of w_{ni_p} and z_{ni_p} are drawn in following two substeps:

a) w_{ni_p} is drawn from the conditional distribution of W_{ni_p} given the data $\mathbf{x}$ and $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$, which is given in (8).

b) z_{ni_p} is drawn from the conditional distribution of Z_{ni_p} given $\mathbf{w}$, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, which is defined as

$$Z_{ni_p} \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{x} \sim \begin{cases} N(\lambda_{ni_p}, 1) \text{ truncated at the left at 0}, & \text{if } w_{ni_p} = 1, \\ N(\lambda_{ni_p}, 1) \text{ truncated at the right at 0}, & \text{if } w_{ni_p} = 0. \end{cases} \tag{10}$$

Step 2

The value of $\boldsymbol{\theta}$ is drawn from the conditional posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{z}$ and $\boldsymbol{\xi}$. The distribution is derived as follows. From the definition of the latent variables Z_{ni_p} it follows that $Z_{ni_p} + b_{i_p} = a_{i_p}\theta_n + \varepsilon_{ni_p}$, with ε_{ni_p} being a normally distributed residual. Because (a_{i_p}, b_{i_p}) is fixed, the equality defines a linear model for the regression of $Z_{ni_p} + b_{i_p}$ on a_{i_p}, with regression coefficient θ_n, which has a normal prior with parameters $\mu = 0$

and $\sigma = 1$. Therefore, the posterior of θ_n is also normal. That is,

$$\theta_n \sim N\left(\frac{\hat{\theta}_n/\nu + \mu/\sigma^2}{1/\nu + 1/\sigma^2}, \frac{1}{(1/\nu + 1/\sigma^2)}\right), \tag{11}$$

where

$$\hat{\theta}_n = \left[\sum_p \sum_{i_p} d_{ni_p} a_{i_p} (z_{ni_p} + b_{i_p})\right] / \left[\sum_p \sum_{i_p} d_{ni_p} a_{i_p}\right]$$

and

$$\nu = 1/\left[\sum_p \sum_{i_p} d_{ni_p} a_{i_p}.\right]$$

Step 3

The vector of random item parameters $\boldsymbol{\xi}_{i_p}$ is partitioned into $\boldsymbol{\delta} \equiv (\boldsymbol{\delta}_{i_p}) \equiv (a_{1_1}, b_{1_1}, ..., a_{i_p}, b_{i_p}, ...)$ and $\mathbf{c} \equiv (c_{1_1}, ..., c_{i_p}, ...)$. Hence, their conditional posterior density factors as $p(\boldsymbol{\xi}_{i_p}|\boldsymbol{\theta}, \mathbf{z}_{i_p}, \boldsymbol{\mu}_{i_p}, \boldsymbol{\Sigma}_{i_p}) = p(\text{logit } c_{i_p}|\boldsymbol{\delta}_{i_p}, \boldsymbol{\theta}, \mathbf{z}_{i_p}, \boldsymbol{\mu}_{c|\delta}, \boldsymbol{\Sigma}_{c|\delta})$ $p(\boldsymbol{\delta}_{i_p}|\boldsymbol{\theta}, \mathbf{z}_{i_p}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, where $\boldsymbol{\mu}_{c|\delta}$ and $\boldsymbol{\Sigma}_{c|\delta}$ are the expectation and variance of logit c_{i_p} conditional on $\boldsymbol{\delta}_{i_p}$. Then the following two substeps are made:

a) The value of δ_{i_p} is drawn from the conditional posterior distribution of the parameters of $\boldsymbol{\delta}$ given $\boldsymbol{\theta}$, $\mathbf{z}_{i_p}$, $\boldsymbol{\mu}_p$, and $\boldsymbol{\Sigma}_p$ The distribution is derived as follows: Parameters $\boldsymbol{\delta}_{i_p}$ can be viewed as coefficients of the regression of $\mathbf{z}_{i_p} \equiv (z_{ni_p})$, on $\mathbf{X} \equiv (\boldsymbol{\theta}, -\mathbf{1})$, with $-\mathbf{1}$ being a column vector with entries -1. So we have $\mathbf{z}_{i_p} = \mathbf{X}\boldsymbol{\delta}_{i_p} + \boldsymbol{\varepsilon}_{i_p}$. Only examinees responding to item i_p are considered here. Further, $\boldsymbol{\delta}_{i_p}$ has a normal prior with mean $\boldsymbol{\mu}_p$ and variance $\boldsymbol{\Sigma}_p$. Define $\widehat{\boldsymbol{\delta}}_{i_p} \equiv (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{z}_{i_p}$, define $\mathbf{d} \equiv \mathbf{X}^t\mathbf{X}\widehat{\boldsymbol{\delta}}_{i_p} + \boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}_p$ and define $\mathbf{D} \equiv \left(\mathbf{X}^t\mathbf{X} + \boldsymbol{\Sigma}_p^{-1}\right)^{-1}$. Then a well-known result from Bayesian regression analysis (see, for instance, Box & Tiao, 1973) is that

$$\boldsymbol{\delta}_{i_p} \mid \boldsymbol{\theta}, \mathbf{z}_{i_p}, \mathbf{X}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p \sim N(\mathbf{Dd}, \mathbf{D}). \tag{12}$$

b) The value of c_{i_p} is sample from the conditional posterior distribution given δ_{i_p}, θ, $\mathbf{z}_{i_p}$, $\mu_{c|\delta}$,and $\Sigma_{c|\delta}$. Let t_{i_p} be the number of persons who do not know the correct answer to item i_p and guess the response. For the probability of a correct response of a person n on item i_p given $w_{ni_p} = 0$ it thus holds that $P(Y_{ni_p} = 1 \mid W_{ni_p} = 0) = c_{i_p}$. The number of correct responses obtained by guessing, S_{i_p}, say, has a binomial distribution with parameters c_{i_p} and t_{i_p}. Since logit c_{i_p} has a normal prior with parameters $\mu_{c|\delta}$ and $\Sigma_{c|\delta}$, the procedure for sampling in a generalized linear model with a logit-link and a normal prior (see, Gelman Carlin, Steam & Hall, 1995, sects 9.9 and 10.6) can be used.

Step 4

Values for (μ_p, Σ_p) are drawn from the conditional posterior distribution given ξ, θ, $\mathbf{z}$, and $\mathbf{x}$. The number of items sampled from population p is equal k_p. The prior distribution in (5) is the conjugate for (μ_p, Σ_p). Hence, the posterior distribution is also normal-inverse-Wishart, with parameters

$$\mu_p = \frac{\kappa_0}{\kappa_p}\mu_0 + \frac{k_p}{\kappa_p}\bar{\xi}_p \tag{13}$$

$$\nu_p\Sigma_p = \nu_0\Sigma_o + \mathbf{S} + \frac{\kappa_0 k_p}{\kappa_p}(\bar{\xi}_p - \mu_0)(\bar{\xi}_p - \mu_0)^t, \tag{14}$$

where $\kappa_p = \kappa_0 + k_p$, $\nu_p = \nu_0 + k_p$, $\mathbf{S} = \sum_{i_p}^{k}(\xi_{i_p} - \bar{\xi}_p)(\xi_{i_p} - \bar{\xi}_p)^t$ and $\bar{\xi}_p = \sum_{i_p}^{k}\xi_{i_p}$. The corresponding posterior distribution is thus given by

$$\begin{aligned} \Sigma_p \mid \xi_p &\sim \text{Inverse-Wishart}_{k-1}(\mathbf{S}^{-1}), \\ \mu_p \mid \Sigma_p, \bar{\xi}_p &\sim N(\bar{\xi}_p, \Sigma_p/k). \end{aligned} \tag{15}$$

The procedure thus amounts to iterative generation of parameter values using the above four steps. Multiple MCMC chains can be started from different points to evaluate convergence by comparing the between- and within-sequence variance. Another approach is to generate a single MCMC chain and to evaluate convergence by dividing

the chain into subchains and comparing between- and within-subchain variance. For these and other technical details, see Gelman, Carlin, Steam and Hall (1995).

Necessary Condition on Sampling Design

As already discussed, the procedure breaks down if the examinees are administered unique items. This point can now be illustrated using the steps in the above Gibbs sampler. For example, Step 3a is based on a normal linear model $\mathbf{z}_{i_p} = \mathbf{X}\boldsymbol{\delta}_{i_p} + \boldsymbol{\epsilon}_{i_p}$. However, if random item i_p is administered to one examinee, $\mathbf{z}_{i_p}$ has only one entry, and it is not possible to estimate two regression coefficients from one observation. Likewise, the generalized linear model in Step 3b is based on one observation, and here the estimation procedure also breaks down. A solution to this problem is marginalization of the likelihood function over the random item parameters. The remaining structural parameters are the hyperparameters (μ_p, Σ_p). The estimation equations for these parameters based on the marginal likelihood are given in the next section (cf. van der Linden & Glas, 2001).

Bayes Modal Estimation

All Item Parameters Structural

In Bayes modal estimation (Bock & Aitkin, 1982), a distinction in made between structural and nuisance parameters. If the number of item parameters is limited, that is, this number does not depend on the number of respondents, the item parameters can be viewed as structural parameters and the ability parameters are the nuisance parameters. The structural parameters are stacked in a vector $\eta = (\xi_{1_1}, ..., \xi_{i_p}, ..., \mu_1, \Sigma_1, ..., \mu_p, \Sigma_p, ..., \mu_P, \Sigma_P)$. The structural parameters are estimated from a log-likelihood marginalized with respect to the nuisance parameters. That is, the so-called complete data likelihood given by (6) is integrated over the nuisance parameters. As a result, the marginal probability of observing response pattern $\mathbf{x}_n$ is given

by

$$p(\mathbf{x}_n \mid \mathbf{d}_n, \boldsymbol{\xi}_p) = \int \prod_{i_p} p(x_{ni_p} \mid d_{ni_p}, \theta_n, \boldsymbol{\xi}_{i_p}) p(\theta_n) d\theta_n,$$

and the marginal log-likelihood function of $\boldsymbol{\eta}$ is given by

$$\log L(\boldsymbol{\eta}; \mathbf{x}) = \sum_p \sum_n \left[\log p(\mathbf{x}_n \mid \mathbf{d}_n, \boldsymbol{\xi}_p) + \sum_{i_p} \log p(\boldsymbol{\xi}_{i_p} \mid d_{ni_p}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \right] \tag{16}$$

$$+ \log p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \tag{17}$$

As above, the convention is used that $p(x_{ni_p} \mid d_{ni_p} = 0, \theta_n, \xi_{i_p}) = 1$. The marginal likelihood equations for $\boldsymbol{\eta}$ can be easily derived using Fisher's identity (Efron, 1977; Louis 1982; see also Glas, 1992, 1998). The first-order derivatives with respect to $\boldsymbol{\eta}$ can be written as

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log L(\boldsymbol{\eta}; \mathbf{x}) = \sum_p \sum_n E\left(\frac{\partial}{\partial \boldsymbol{\eta}} \log f_{p,n}(\boldsymbol{\eta}, \boldsymbol{\theta}_n; \mathbf{x}_n) \mid \mathbf{x}_n, \boldsymbol{\eta}\right) = \mathbf{0}, \tag{18}$$

where $\sum_p \sum_n \log f_{p,n}(\boldsymbol{\eta}, \boldsymbol{\theta}_n)$ is the complete data log-likelihood, that is,

$$\sum_p \sum_n \log f_{p,n}(\boldsymbol{\eta}, \boldsymbol{\theta}_n; \mathbf{x}_n) =$$

$$\sum_p \sum_n \log \left[p(\mathbf{x}_n \mid \mathbf{d}_n, \boldsymbol{\xi}_p, \theta_n) + \log p(\theta_n) + \sum_{i_p} \log p(\boldsymbol{\xi}_{i_p} \mid d_{ni_p}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \right]$$

$$+ \log p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

Notice that the first-order derivatives in (18) are expectations with respect to the conditional posterior density of the nuisance parameters.

Let P_{ni_p} and Φ_{ni_p} be defined by $P_{ni_p} = p(x_{ni_p} \mid d_{ni_p} = 1, \theta_n, \xi_{i_p}) = c_{i_p} + (1 - c_{i_p})\Phi_{ni_p}$, so Φ_{ni_p} is the normal-ogive part of the probability P_{ni_p}. By taking first order derivatives of the logarithm of this expression, likelihood equations for the parameters

$\xi_{i_p u}, u = 1, ..., 3$, are found as

$$\sum_{n|d_{ni_p}=1} E\left(\frac{(x_{ni_p} - P_{ni_p})\Phi_{ni_p}(\theta_n - b_{i_p})}{P_{ni_p}} \middle| \mathbf{x}_n, \boldsymbol{\eta} \right) + (a_{i_p} - \mu_{p1}) = 0, \qquad (19)$$

$$\sum_{n|d_{ni_p}=1} E\left(\frac{(P_{ni_p} - x_{ni_p})\Phi_{ni_p} a_{i_p}}{P_{ni_p}} \middle| \mathbf{x}_n, \boldsymbol{\eta} \right) + (b_{i_p} - \mu_{p2}) = 0, \qquad (20)$$

and

$$\sum_{n|d_{ni_p}=1} E\left(\frac{(x_{ni_p} - P_{ni_p})(1 - \Phi_{ni_p})c_{i_p}(1 - c_{i_p})}{P_{ni_p}(1 - P_{ni_p})} \middle| \mathbf{x}_n, \boldsymbol{\eta} \right) + (\text{logit}c_{i_p} - \mu_{p3}) = 0. \quad (21)$$

These expressions are a straightforward generalization of the usual likelihood equations for the 3PNO; for details, refer to Glas (2000). It is easily verified that the likelihood equation for the parameters of the parent items are given by (13) and (14). The likelihood equations be solved using an EM or Newton-Raphson algorithm. Since the number of parameters in practical applications will be quite large, the latter algorithm will seldom be feasible. Expressions for confidence intervals can also be derived using Fisher's identity (Louis 1982; Mislevy, 1986, Glas, 1998). However, the computation of the asymptotic covariance matrix of the estimates also involves the inversion of a matrix of second-order derivatives (information matrix). In the application presented below, only the information matrix within the item populations will be inverted, that is, the covariance between the populations will be assumed zero. This approximation results in confidence intervals that are larger than the confidence intervals obtained when the complete information matrix would be inverted.

Incidental Item Parameters

In the case every person is administered a unique item, say in a procedure where items are generated on the fly, the situation is different in the sense that the random item

parameters are unique for every person. This will be made explicit by adding an index n and writing $\boldsymbol{\xi}_{i_p n}$. Now the number of item parameters $\boldsymbol{\xi}_{i_p n}$ grows with the sample size, so it is doubtable whether they can be consistently estimated, and, therefore, they must be viewed as nuisance parameters, together with the ability parameters (Neyman & Scott, 1948; Kiefer & Wolfowitz, 1956). These nuisance parameters are stacked in vectors $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, respectively. This leaves $\boldsymbol{\eta} = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p, ..., \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ as structural parameters. The marginal probability of observing response pattern $\mathbf{x}_n$ is now given by

$$p(\mathbf{x}_n; \boldsymbol{\eta}) = \int \cdots \int \prod_{p,i_p} p(x_{ni_p} \mid d_{ni_p}, \theta_n, \boldsymbol{\xi}_{i_p n}) p(\boldsymbol{\xi}_{i_p n} | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) p(\theta_n) d\boldsymbol{\xi}_{i_p n} d\theta_n$$

$$= \int \left[\prod_{p,i_p} \int \cdots \int p(x_{ni_p} \mid d_{ni_p}, \theta_n, \boldsymbol{\xi}_{i_p n}) p(\boldsymbol{\xi}_{i_p n} | d_{ni_p}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) d\boldsymbol{\xi}_{i_p n} \right] p(\theta_n) d\theta_n.$$

Notice that (8) entails a multiple integral over $\boldsymbol{\xi}_{i_p n}$. It now follows that the likelihood equations are given by

$$\mu_{pu} = \frac{1}{N} \sum_n E(\xi_{pu} \mid \mathbf{x}_n, \boldsymbol{\eta}), \tag{22}$$

$$\sigma^2_{pu} = \frac{1}{N} \sum_n E(\xi^2_{pu} \mid \mathbf{x}_n, \boldsymbol{\eta}) - \mu^2_{pu}, \tag{23}$$

and

$$\sigma_{puv} = \frac{1}{N} \sum_n E(\xi_{pu} \xi_{pv} \mid \mathbf{x}_n, \boldsymbol{\eta}) - \mu_{pu} \mu_{pv}, \tag{24}$$

where indices u and $v \neq u$ denote the uth and vth element in the parameter vectors. Again, these equations can be solved using an EM or Newton-Raphson algorithm.

Discussion

A final remark concerns a case where each generated item is administered to more than one person, but the number of generated items still grows with the sample size. In

that case, the random item parameters must still be considered as nuisance parameters. Consider the case where each random item is given to two respondents, say n and m. The responses of both respondents now depend on the same random item parameter; this dependency will be made explicit by labeling this item parameter as $\xi_{i_p nm}$. The complete-data likelihood can now be written as

$$p(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{(n,m)} \prod_{p} \prod_{i_p} p(x_{ni_p} \mid d_{ni_p}, \theta_n, \xi_{i_p nm})$$
$$p(x_{mi_p} \mid d_{mi_p}, \theta_m, \xi_{i_p nm}) p(\xi_{i_p nm} \mid d_{mi_p}, \mu_p, \Sigma_p) p(\theta_n) p(\theta_m),$$

where the product is over pairs of respondents, and marginalized results in

$$p(\mathbf{x}; \eta) = \prod_{(n,m)} \int \int \prod_{p} \left[\int p(x_{ni_p} \mid d_{ni_p}, \theta_n, \xi_{i_p nm}) \right.$$
$$\left. p(x_{mi_p} \mid d_{mi_p}, \theta_m, \xi_{i_p nm}) p(\xi_{i_p nm} | d_{mi_p}, \mu_p, \Sigma_p) d\xi_{i_p n} \right] p(\theta_n) p(\theta_m) d\theta_n d\theta_m$$

Notice that the integral does not factor further. In fact, as the number of respondents receiving the same random item goes up, we are quickly left with a multiple integral that cannot be computed by the usual Gauss-Hermite procedure (see, for instance, Glas, 1992). Fortunately, the fully Bayesian procedure discussed above does not have these problems.

Some Numerical Examples

A number of studies were conducted to assess the feasibility of the procedures in practical situations. In some practical situations, the number of responses per population of items might be quite low and the number of item parameters might be quite high. In such cases, the convergence of the MCMC- or the EM-algorithm to reasonable parameter estimates is not a priori obvious. On the other hand, in a Bayesian framework the computation of estimates can be supported by a sensible choice of priors.

The study consisted of two stages. In the first stage, two real data sets were analyzed to obtain some idea of the covariance between the item parameters. Then, in the second stage, the estimates obtained in the first stage were used in a number of simulation studies aimed at assessing the quality of parameter recovery.

The first data set consisted of the responses of 429 students to 10 multiple choice items in a computer based test for a course on naval architecture at the Ngee Ann Polytechnic in Singapore. The data were collected in 1999 and 2000. The numerical information in the item stem and the response alternatives was randomly changed every time an item was administered. The second data set consisted of the responses of a sample of 4000 students from the population participating in the 1991 central examination on French language comprehension in Secondary Education in the Netherlands. In this case, the test was a traditional paper-and-pencil test. Students were clustered in 116 schools, and it was assumed that the item parameters varied across classes. It was expected that the item-parameter variance might be high in the first example and low in the second example. In the first example, Bayes modal estimates of μ and Σ were obtained by marginalizing over all incidental parameters ξ and θ. In the second example, two procedures were used. In the first procedure, concurrent estimates of ξ, θ, μ and Σ were obtained using the MCMC method run with 13,000 iterations, 3000 of which were burn-in iterations. Below, expected a posteriori (EAP) estimates are reported as point estimates. In the second procedure, Bayes modal (MAP) estimates of ξ, μ and Σ were obtained by marginalizing over θ. Computations were carried out using the EM-algorithm.

In both examples, the same prior covariance matrix Σ_0 was used. The values in Σ_0 are shown in Table 1; they are no more than an educated guess. For instance, the negative covariance between the discrimination parameter a_{i_p} and the logit-guessing-parameter logit c_{i_p} is based on the consideration that, to obtain similar the item characteristic curves (ICCs), the discrimination parameter must go down if the lower asymptote goes up. In the same manner, when the respondents are relatively proficient, lowering the item difficulty parameter can be counterbalanced by lowering the discrimination parameter. This feature accounts for the choice of a positive prior covariance between the two parameters. The

prior for the parent item parameters was chosen equal to μ_0= (1.0, 0.0,logit(0.25)). To obtain convergence in the analysis of the language comprehension data, it turned out that the parameters in the normal-inverse-Wishart prior for (μ_p, Σ_p) had to be set equal to $\nu_0 = 10$ and $\kappa_0 = 10$, respectively. Since $k_p = 160$, this choice results in a slightly informative prior. An uninformative prior sufficed for the Naval Architecture data.

The averages of the point estimates of the covariance matrices are shown in Table 1 (first three columns), together with their confidence intervals (last three columns). It can be seen that both the posterior variance of the item discrimination and difficulty parameters was generally lower than expected. For the EAP estimates, the posterior standard deviation is reported; for the MAP estimates, the values computed using the normal approximation are shown. It can be seen that the estimated variances are lower than the prior variances. Further, the standard errors of the MAP estimates are smaller than those of the EAP estimates. This effect is consistent with the findings of Glas, Wainer and Bradlow (2000). They argue that posterior distributions of bounded parameters, such as a variance or a discrimination parameter, are skewed. The standard error of the MAP estimate used here is based on an assumption of asymptotic normality, which, in turn, is based on a Taylor-expansion of the likelihood which terms of order higher two ignored. The fact that here only the within-item-population information matrices were used to obtain the standard errors did not nullify the effect.

[Table 1 about here]

The second part of this study was aimed at assessing the quality of parameter recovery. Since the difference in the covariances obtained for the two examples given above was not dramatically different, it was decided to study two conditions in two simulation studies. In the first simulation study, the prior parameters μ_0 and Σ_0 were the same as in the examples presented above. The parent item parameters, μ_p were drawn from a normal distribution indexed by μ_0 and Σ_0, and Σ_p was set equal to Σ_p. Then, for each population, 10 items were randomly drawn from a normal distribution with parameters μ_p and Σ_p. To produce realistic data, parent and random item discrimination parameters drawn below 0.5 were truncated to 0.5. The responses to the random items

were generated for simulees with an ability parameter randomly drawn from a standard normal distribution. Every simulee responded to 20 random items from 20 different populations. So the total data matrix consisted of 1,000 responses. As above, 13,000 iterations were made, including 3,000 burn-in iterations. To obtain convergence, the parameters in the normal-inverse-Wishart prior had to be set equal to $\nu_0 = 2$ and $\kappa_0 = 2$, respectively. Since $k_p = 10$, this choice entails a quite informative prior.

The second simulation study had a similar set-up. The average of the EAP estimates of the mean and covariance matrix obtained using the French language examination was used as μ_0 and Σ_0. Further, the number of item populations was equal to 40, the number of random items per population was qual to 20, and the number of responses to each random item was 200. So in this case, the total number of responses was equal to 4,000.

[Table 2 about here]

Some results of the two simulations are presented Table 2. The results are averaged over 10 replications and all items. The rows labeled a, b and logit c relate to random item parameters; all other rows relate to item-population parameters. The two columns labeled EAP relate to EAP estimates obtained using the Gibbs sampler, the columns labeled MAP relate to Bayes modal estimates from a posterior marginalized over θ. The columns labeled MAE give the mean absolute error of the estimates, averaged over items and replications. The columns labeled SE give the posterior standard deviation and the normal approximation for the EAP and the MAP estimates, respectively, again averaged over items and replications. It can be seen that the magnitudes of these estimates are clearly smaller than the corresponding MAEs. Further, especially in the case $P = 40$, the estimates of the covariance matrices seemed much more precise than the estimates of the item parameters. This result, however, is explained by the fact that the covariance matrices were not varied over populations, but chosen equal to their prior values. Further inspection of the results shows that the MAEs of the MAP estimates were somewhat smaller than the corresponding EAP estimates.

[Figure 1 about here]

Figure 1 shows the posterior distributions of a typical set of parameters for a run with $P = 20$. The three pictures in the first row are the posterior distributions of the three elements of μ_p for a typical item-population p. The three pictures in the next row show the posterior distributions of the three parameters of an arbitrarily chosen random item i_p. The last two rows give the posterior distributions of the elements of Σ_p, for the same item-population p. The dotted line in the pictures are the asymptotic distributions computed using the normal approximation described above. It can be seen that the latter approximations are not always realistic. The normal approximation of the variance of logit c_{i_p}, for instance, gives discernible larger positive weight to negative values. The actual posterior distributions of several elements of Σ_p are notably skewed to the right. Figure 2 shows the convergence of the Gibbs sampler for the same 12 parameters. The plot is based on the 2,000 draws taken equally spaced from the 10,000 draws following the burn-in iterations. From inspection of the plots it can be concluded that the chain has properly converged. In practice, visual inspection of the convergence plots of all parameters is not very practical. However, convergence can also be evaluated by dividing the generated chain into batches and comparing the within and between batch variance of the generated values.

[Figure 2 about here]

Finally, the figures 3 and 4 give a scatter plot of the generating values (x-axis) and the EAP-estimates (y-axis) of the population and random item parameters for two replications of both simulation studies. The truncation of the discrimination parameters at 0.5 is caused by the generation strategy described above. It can be seen from the plots that the relation between the generated and recovered parameters is quite good; in fact, all four correlations were above 0.80. Similar plots could not be made for logit c_{i_p} and its mean and the elements of the covariance matrices, because the variance in the generating values was too low and zero, respectively.

Discussion

Several authors (Bradlow, Wainer & Wang, 1999; Janssen, Tuerlinckx, Meulders & de Boeck, 2000; Wainer, Bradlow & Du, 2001) have proposed IRT models with random item parameters. These models, however, do not include within-item covariance of random item parameters. In this article, such a model was proposed, and Bayesian estimation methods for such models were outlined. It was shown that the sampling design is a crucial factor here. If every random item is responded to by a substantial number of respondents, Bayesian methods using the Gibbs sampler or marginalization over the ability parameters can be used. If only one response is given to every random item, these approaches break down. However, in that case, and only then, a Bayes modal estimation procedure using a posterior distribution marginalized with respect to ability parameters and the random item parameters can be used to estimate the means and covariance matrices of the item population parameters.

A rule of thump for the minimum number of respondents that should respond to a random item in the first case was not sought here. However, already with 10 and 20 random items per parent and 100 and 200 responses to every random item, the prior on the covariance matrix had to be informative. Situations with fewer random item parameters and observations per random item parameter might be modeled by assuming that all item parents have the same covariance matrix, but this suggestion remains a point of further study.

References

Albert, J.H. (1992). Bayesian estimation of normal-ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics, 17* 261-269.

Albers, W., Does, R. J. M. M., Imbos, Tj., & Janssen, M. P. E. (1989). A stochastic growth model applied to repeated test of academic knowledge. *Psychometrika, 54*, 451-466.

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika, 42,* 357-374.

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika, 66.*

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-357). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bock, R.D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scores items. *Psychometrika, 35,* 179-197.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika, 46,* 443-459.

Box, G. and Tiao, G. (1973). *Bayesian inference in statistical analysis.* Addison-Wesley Publishing Company, Reading, Massachusetts.

Bradlow, E. T., Wainer, H., Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Efron, B. (1977). Discussion on maximum likelihood from incomplete data via the EM algorithm (by A. P. Demster, N. M. Laird and D. B. Rubin). *Journal of the Royal Statistical Society* (Series B), *39,* 1-38.

Fox, J.P., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271-288.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85,* 398-409.

Gelman, A, Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian data analysis.*

London: Chapman and Hall.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1; pp. 236-258). Norwood, NJ: Ablex Publishing Corporation.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8,* 647-667.

Glas, C. A. W. (2000). Item calibration and parameter drift. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 183-199). Norwell, MA: Kluwer Academic Publishers.

Glas, C. A. W., & van der Linden, W. J. (2001). *Computerized adaptive testing with item clones*. Submitted for publication.

Glas, C.A.W., Wainer, H., & Bradlow (2000). MML and EAP estimates for the testlet response model. In W.J. van der Linden & C.A.W.Glas (Eds.), *Computer Adaptive Testing: Theory and Practice* (pp.271-287). Boston MA: Kluwer-Nijhoff Publishing.

Janssen, R., Tuerlinckx, F., Meulders, M. & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285-306.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887-903.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44,* 226-233.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addision-Wesley.

Millman, J. (1973). Passing score and test lengths for domain-referenced measures. *Review of Educational Research, 43,* 205-216.

Millman, J., & Westman, R.S. (1989). Computer-assisted writing of achievement test items: Toward a future technology. *Journal of Educational Measurement, 26,* 177-190.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Pychome-*

trika, 51, 177-195.

Mislevy, R. J. (1991). Randomization-based inferences about latent variables form complex samples. *Pychometrika, 56*, 177-196.

Neyman, J., & Scott, E.L. (1948). Consistent estimates, based on partially consistent observations. *Econometrica, 16*, 1-32.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple Item Types, Missing Data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

Roid, G., & Haladyna, T. (1982). *A technology for test-item writing*. New York: Academic Press.

Sanathanan, L., & Blumenthal, W. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association, 73*, 684-693.

Shi, J. Q., & Lee, S. Y. (1998). Bayesian sampling based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology, 51*, 233-252.

Tsutakawa, R. K., & Johnson, C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 62*, 371-390.

van der Linden, W. J., & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-25). Norwell, MA: Kluwer Academic Publishers.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-269). Norwell, MA: Kluwer Academic Publishers.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-202.

Table 1
Prior and posterior values
item covariance matrix

Prior Covariance Matrix					
.200					
.100	1.000				
-.050	.050	.100			
French Language Comprehension EAP Estimate			SE		
.102			.017		
.031	.208		.017	.033	
-.018	.010	.116	.018	.020	.039
French Language Comprehension MAP Estimate			SE		
.098			.014		
.029	.199		.012	.025	
-.018	.006	.107	.015	.016	.037
Naval Architecture EAP Estimate			SE		
.120			.032		
.027	.122		.030	.051	
.001	.002	.110	.022	.023	.073

Table 2
Parameter Recovery

					EAP		MAP	
P	k_p	n_p	Parameter	True	MAE	SE	MAE	SE
20	10	100	a	1.000	0.404	0.334	0.407	0.330
			b	0.000	0.514	0.346	0.425	0.214
			logit c	-1.099	0.327	0.654	0.322	0.639
			μ_a	1.000	0.311	0.199	0.283	0.107
			μ_b	0.000	0.494	0.307	0.368	0.178
			$\mu_{logit\ c}$	-1.099	0.214	0.414	0.188	0.124
			σ^2_a	0.200	0.076	0.235	0.091	0.059
			σ^2_b	1.000	0.289	0.684	0.080	0.020
			$\sigma^2_{logit\ c}$	0.100	0.377	0.550	0.477	0.108
			$\sigma_{a,b}$	0.100	0.089	0.289	0.067	0.014
			$\sigma_{a,logit\ c}$	-0.050	0.046	0.227	0.051	0.034
			$\sigma_{b,logit\ c}$	0.050	0.071	0.470	0.091	0.055
40	20	200	a	0.950	0.392	0.204	0.424	0.203
			b	0.190	0.365	0.192	0.327	0.141
			logit c	-0.979	0.306	0.294	0.252	0.333
			μ_a	0.960	0.298	0.095	0.261	0.062
			μ_b	0.180	0.318	0.124	0.200	0.076
			$\mu_{logit\ c}$	-1.002	0.199	0.130	0.163	0.104
			σ^2_a	0.102	0.044	0.065	0.040	0.058
			σ^2_b	0.208	0.100	0.118	0.076	0.106
			$\sigma^2_{logit\ c}$	0.116	0.011	0.050	0.014	0.047
			$\sigma_{a,b}$	0.031	0.037	0.066	0.025	0.057
			$\sigma_{a,logit\ c}$	-0.018	0.009	0.043	0.009	0.043
			$\sigma_{b,logit\ c}$	0.010	0.016	0.055	0.016	0.045

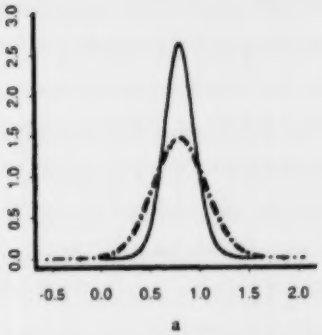
Figure Captions

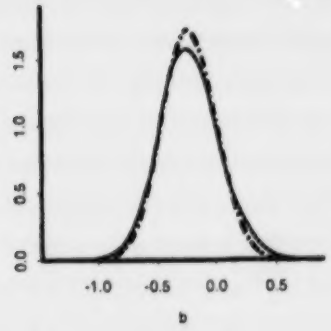
Figure 1. Posterior densities and normal approximations

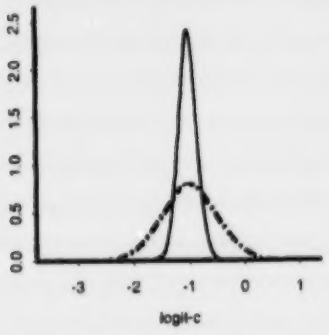
Figure 2. Convergence of the Gibbs samper

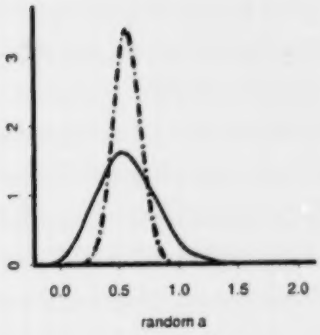
Figure 3. Generating values and parameter estimates $K = 20, k_{i_p} = 10$

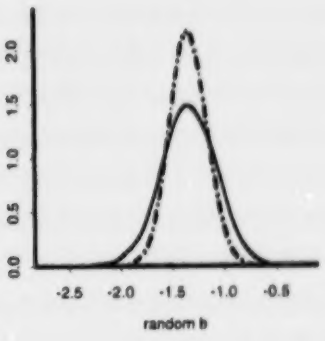
Figure 4. Generating values and parameter estimates $K = 40, k_{i_p} = 20$

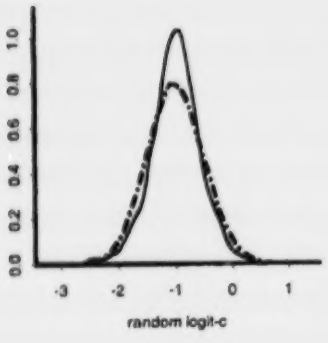


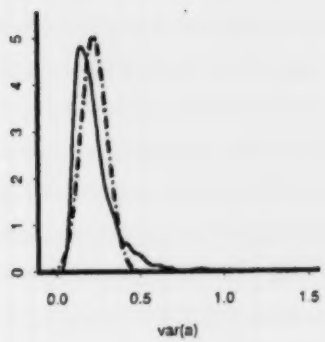


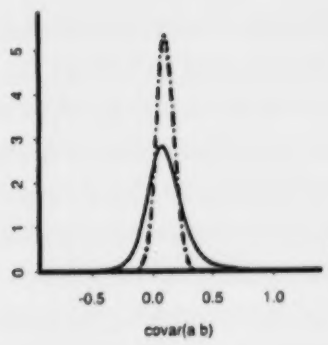


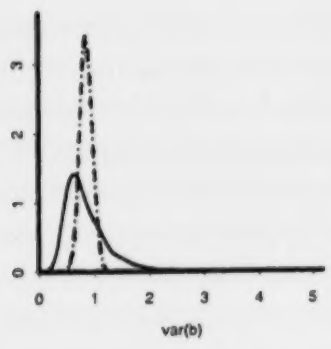


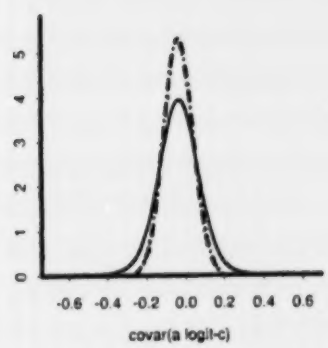


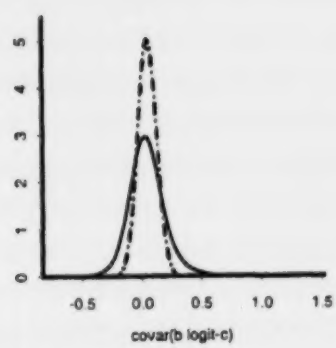


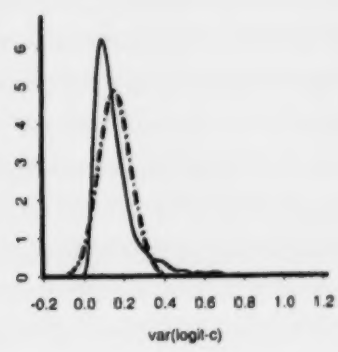




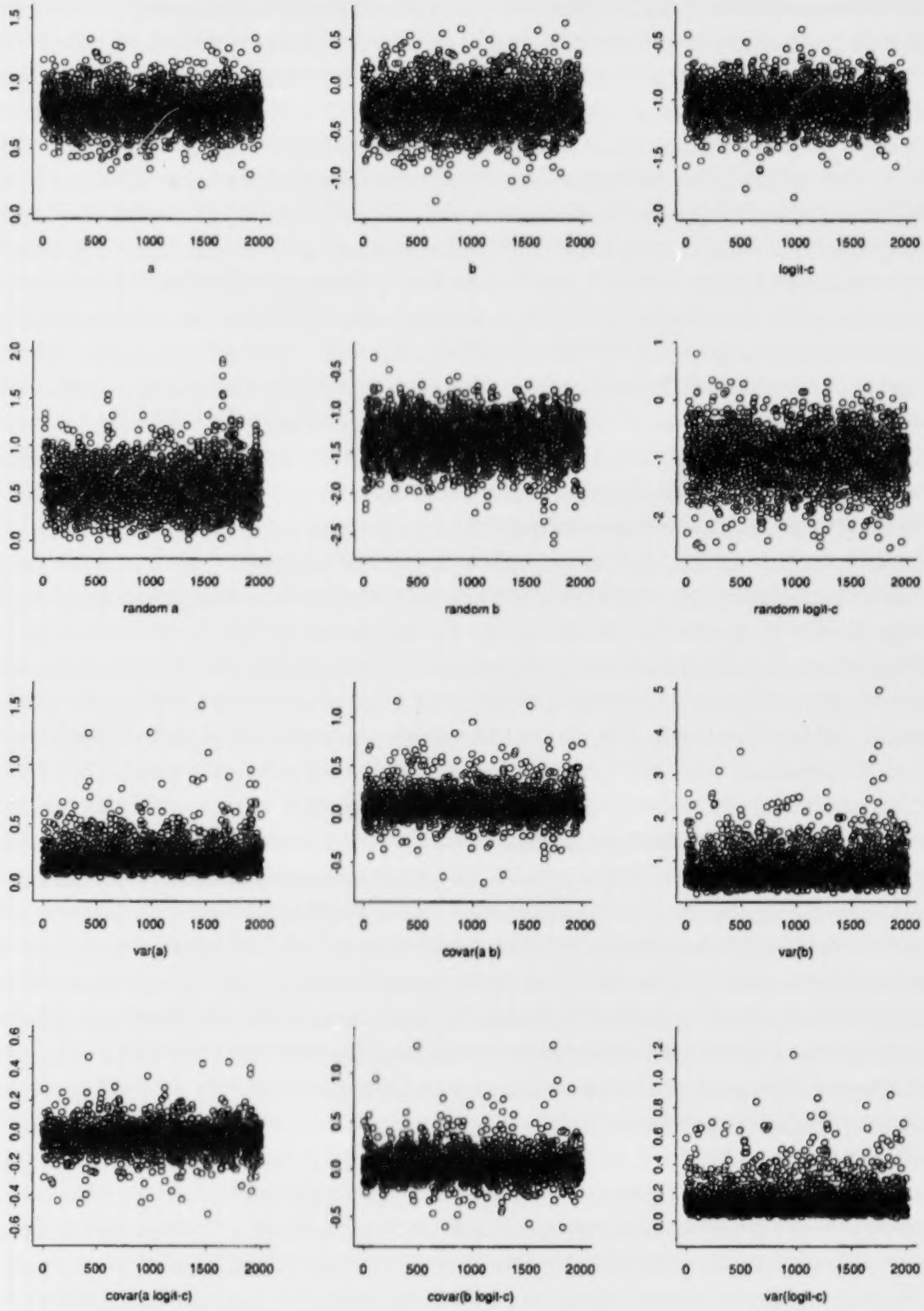




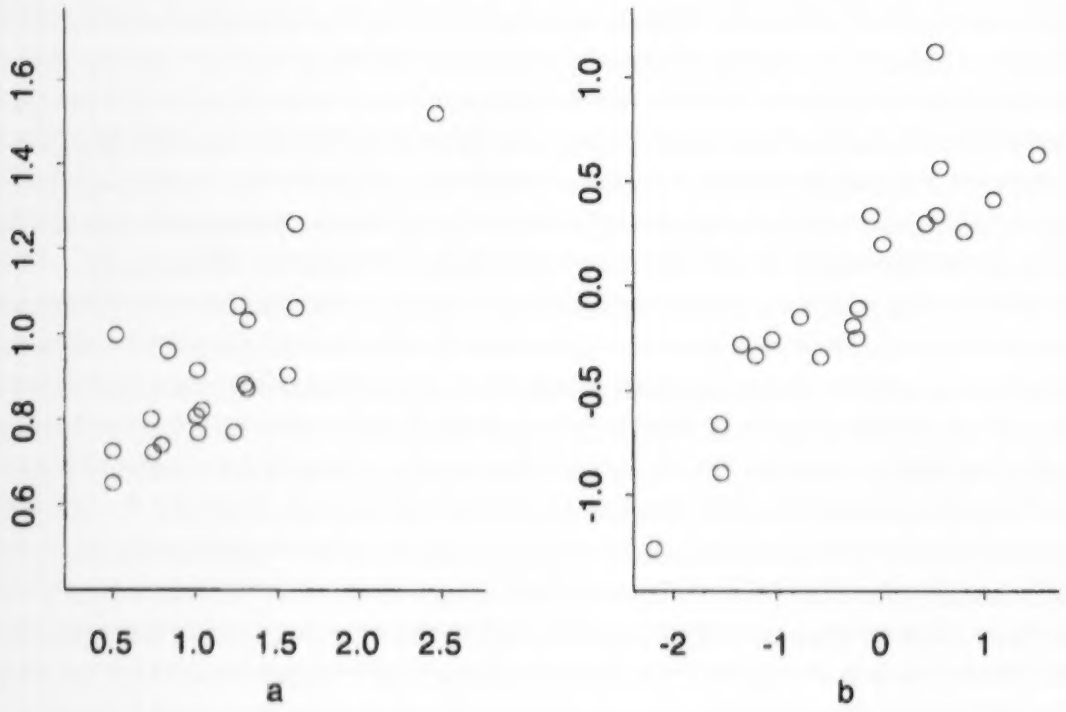


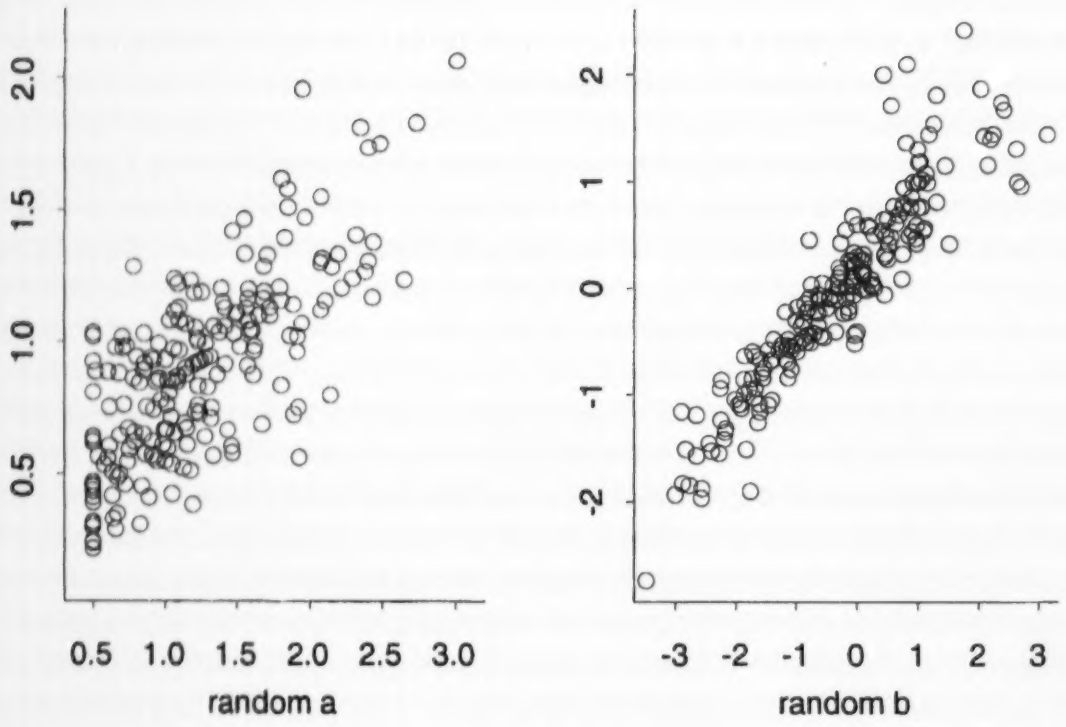


BEST COPY AVAILABLE

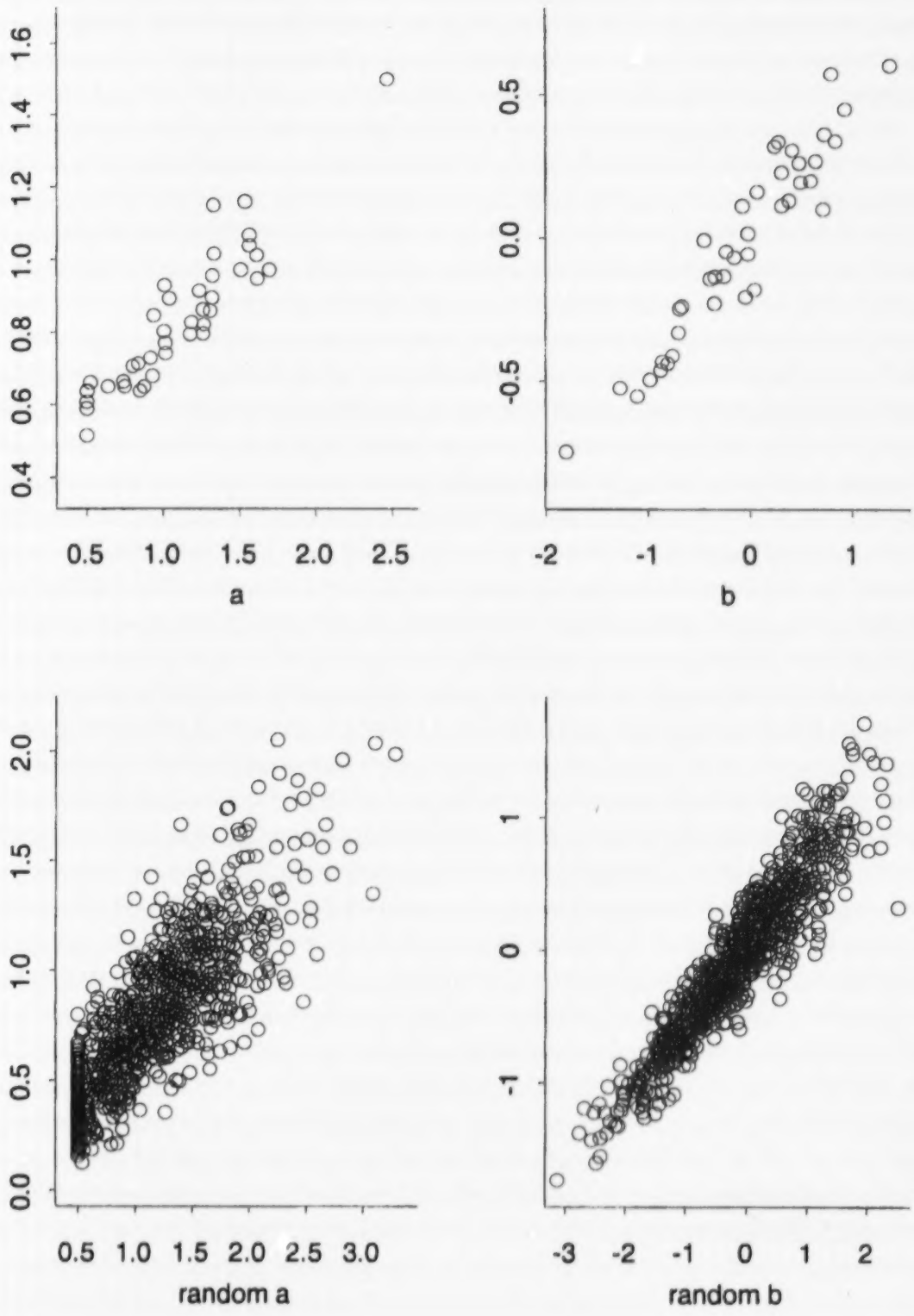


BEST COPY AVAILABLE





BEST COPY AVAILABLE



BEST COPY AVAILABLE

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-01-11 C.A.W. Glas & W.J. van der Linden, *Modeling Variability in Item Parameters in Item Response Models*

RR-01-10 C.A.W. Glas & W.J. van der Linden, *Computerized Adaptive Testing with Item Clones*

RR-01-09 C.A.W. Glas & R.R. Meijer, *A Bayesian Approach to Person Fit Analysis in Item Response Theory Models*

RR-01-08 W.J. van der Linden, *Computerized Test Construction*

RR-01-07 R.R. Meijer & L.S. Sotaridona, *Two New Statistics to Detect Answer Copying*

RR-01-06 R.R. Meijer & L.S. Sotaridona, *Statistical Properties of the K-index for Detecting Answer Copying*

RR-01-05 C.A.W. Glas, I. Hendrawan & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*

RR-01-04 R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*

RR-01-03 R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*

RR-01-02 R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*

RR-01-01 H. Chang & W.J. van der Linden, *Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach*

RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*

RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score Equating*

RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*

RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*

RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").